

Calibration of car ownership models for population group analysis

Erwin Walraven^{1*}, Jorane Rogier¹ and Maaïke Snelder^{1,2}

^{1*}Netherlands Organisation for Applied Scientific Research, The Hague,
The Netherlands.

²Delft University of Technology, Delft, The Netherlands.

*Corresponding author(s). E-mail(s): erwin.walraven@tno.nl;
Contributing authors: jorane.rogier@tno.nl; maaike.snelder@tno.nl;

Abstract

Predicting car ownership of households is a key step in transport modeling, traditionally performed using discrete choice models such as the multinomial logit model, while alternative machine learning algorithms have emerged in recent years. Existing evaluation of prediction models typically focuses on household-level metrics, such as log loss and probability calibration, implicitly assuming that accurate household-level predictions also lead to reliable predictions for population groups. In this paper we demonstrate that this assumption does not necessarily hold. Models that provide well-calibrated probabilities at the household level can still produce inaccurate probability distributions when predictions are aggregated over groups. Motivated by this observation, we identify subgroup-level calibration as a distinct prediction objective, and propose a model-agnostic calibration method that explicitly improves the accuracy of predicted distributions for population groups. In a case study based on a large dataset of Dutch households, we evaluate multinomial logit, ordered logit, random forest and gradient boosting models. The results show that subgroup-level prediction errors can remain even when household-level calibration metrics indicate good performance, while the proposed post-processing approach improves subgroup-level predictions across different model classes. These findings highlight the importance of evaluating prediction models at the level at which they are ultimately used, which is particularly relevant for transport policy applications involving equity and population-group analysis.

Keywords: car ownership prediction, population groups, probability calibration, machine learning

1 Introduction

The continuous growth of cities puts increasing pressure on urban areas, in which parts of the mobility system such as public transport, parking, mobility hubs, and road networks occupy scarce urban space (Mao et al. 2020). Due to the scarcity of space, there is an increasing focus on understanding how interventions influence travel behavior, and how such behavioral changes can potentially affect the use of space in a city. An example is the introduction of car-low-areas (Jorritsma et al. 2023), in which parking places may be removed if travelers shift to other modes of transport and consequently own fewer cars. The space that becomes available can be used for other activities such as leisure and sports. When creating car-low areas with fewer parking places, it is crucial to understand which factors influence the decision of households to own a car, how accessibility can be preserved by offering alternative travel options, and how this varies across population groups.

Predicting car ownership therefore is a key step in transport planning when adjusting the mobility system and its use of space in an area. First, car ownership models are used in transport models to determine the number of cars owned by individual households in a population, in such a way that it can be taken into account in subsequent modeling steps such as mode choice. Second, car ownership predictions are used to analyze the differences in car ownership between various population groups, depending on characteristics such as age and income. These two uses require related but different forms of predictive reliability. For household-level applications it is important that the predicted probability distribution over the number of cars is meaningful for individual households. For population-group analysis it is important that the aggregation of these household-level predictions leads to accurate distributions for the groups being studied. The distinction is important because transport policy questions are often answered at the level of population groups rather than at the level of individual households.

Traditionally, discrete choice models have been used in the literature to predict and analyze travel behavior, such as the multinomial logit model and its extensions (Li 2011). In recent years alternative machine learning techniques have been proposed, such as decision trees (Kashifi et al. 2022), random forests (Cheng et al. 2019) and neural networks (van Cranenburgh and Alwosheel 2019). These techniques are often presented as competitive or even better than discrete choice models (Paredes et al. 2017). At the same time it is observed that existing research shows mixed or inconsistent results, because in some cases it turns out that a multinomial logit model still performs better (Ali et al. 2023). An important reason why such comparisons are difficult to interpret is that the suitability of a model depends on the use case and on the metrics used to evaluate it. A model may perform well in terms of accuracy, precision,

recall, log loss, or household-level calibration, while still being unsuitable for applications that require reliable probability distributions after aggregation over population groups.

This issue is particularly relevant because distributional effects across population groups are becoming increasingly important in transport planning. In addition to assessing the overall impact of interventions, policy makers often need to understand whether specific groups benefit, are disadvantaged, or respond differently to a proposed measure. Such questions are central to equity analysis and to the design of interventions that do not disproportionately affect parts of the population. This creates a demand for prediction models that provide reliable outputs at the level at which policy questions are asked. For car ownership prediction, this means that a model should not only provide plausible probabilities for individual households, but should also produce accurate car ownership distributions for population groups.

Most existing evaluation and calibration approaches focus on individual predictions. For probabilistic classifiers, metrics such as log loss, reliability diagrams and expected calibration error assess whether predicted probabilities are consistent with observed outcomes at the level of individual observations. Standard calibration techniques (Silva Filho et al. 2023) have been designed with this objective in mind. However, good calibration at the household level does not necessarily imply that predicted distributions are accurate after aggregation over subgroups of the population. A model can assign probabilities that are well calibrated on average across individual households, while still overestimating or underestimating the share of households with zero, one, two, or more cars within specific population groups. This means that household-level calibration and subgroup-level calibration should be treated as distinct prediction objectives.

In this paper, we focus on this distinction. We investigate car ownership prediction models for applications in which predicted probability distributions are used at both the household level and the level of population groups. Our first contribution is to demonstrate that household-level calibration does not necessarily imply accurate predictions at subgroup level. This is a non-trivial observation because prediction models are commonly evaluated using household-level metrics, while many transport policy applications rely on aggregated predictions for groups. Our second contribution is to identify subgroup-level calibration as a distinct prediction objective. Rather than asking only whether predicted probabilities are reliable for households, we explicitly consider whether the average predicted distribution within a group matches the observed distribution for that group. Our third contribution is a model-agnostic calibration method that adjusts the predicted probabilities as a post-processing step in order to improve group-level predictions. The method operates directly on predicted probabilities, and therefore it can be applied to different model classes without modifying their estimation procedures.

We demonstrate these ideas in a case study based on a real-world dataset containing information about a large number of Dutch households. The analysis includes multinomial logit, ordered logit, random forest, and gradient boosting models. This allows us to study whether subgroup-level prediction errors occur across different types

of models, including models that are specifically designed to provide probabilistic outputs. The results show that subgroup-level errors can remain substantial even when household-level calibration metrics suggest good performance. They also show that the proposed post-processing method can improve subgroup-level predictions across different model classes, although the magnitude of the improvement depends on the underlying model and the groups considered. The main implication is that prediction models should be evaluated at the level at which they are ultimately used. For transport policy applications involving population-group analysis, this means that household-level metrics alone are insufficient.

It is important to further clarify both the objective of this work and the notion of calibration considered. The primary goal of this paper is to improve predictive accuracy of probability distributions at the level of population groups, rather than to enable behavioral interpretation. Because the proposed method operates as a post-processing adjustment, it does not preserve the utility-based meaning of coefficients in discrete choice models and should therefore not be used to derive behavioral parameters, elasticities, or welfare measures. Instead, we focus on applied, data-driven prediction tasks where accurate distributional outputs are required. In this context, we adopt the notion of probability calibration from the machine learning literature, in which calibration refers to the consistency between predicted probabilities and observed outcomes. This differs from the traditional use of calibration in transport modeling, where model outputs are adjusted to match external control totals, such as traffic counts, to correct for sampling bias or ensure consistency with the population. The focus of this paper is not to replace model specification or behavioral modeling, but to show that subgroup-level prediction accuracy is a separate objective that requires explicit evaluation and dedicated correction.

The structure of the paper is as follows. In Section 2 we describe background information on car ownership prediction models, relevant metrics and probability calibration. In Section 3 we formalize subgroup-level calibration and propose a calibration method specifically for population groups. Section 4 describes a case study based on data and an evaluation of the prediction models, followed by related work and conclusions in Section 5 and 6.

2 Background on car ownership prediction

In this section we formally introduce car ownership prediction including the mathematical notation that we use throughout this paper. Furthermore, we distinguish between household-level and population-group-level prediction tasks, introduce the corresponding evaluation metrics, and discuss why good performance on one task does not necessarily imply good performance on the other.

2.1 Predictions for households

In this paper we distinguish between predictions that are used at the level of individual households and predictions that are used after aggregating households into population groups. We consider households $h \in H$, and each such household owns at most n cars. Households are characterized by m household properties, such as income and age, and

h_j denotes property j of household h for $j = 1, \dots, m$. A prediction model $P : H \rightarrow [0, 1]^{n+1}$ takes household properties and predicts a probability distribution over the number of cars. The model P can be any prediction model and is estimated based on data.

Multiple types of prediction tasks can be performed by using a prediction model P , each of which requires different performance metrics. In macroscopic transport models it is typically required to assign a fixed number of cars to a household h , which can be done by taking the number of cars for which the predicted score in the vector P_h is maximum. When fitting the model on data, its parameters can be chosen in such a way that the model maximizes common machine learning metrics such as the precision, recall or F-score (Rainio et al. 2024). For models that are not trained with a probabilistic objective, the individual elements of the vector P_h cannot necessarily be interpreted as well-calibrated probabilities. Even when models are trained with probabilistic objectives and provide well-calibrated predictions at the individual level, this does not guarantee that aggregated predictions over population groups are accurate.

In other applications it is required to use the entire distribution P_h that is predicted for a given household h . If one asks “what is the probability that household h owns two cars?”, the value corresponding to this number of cars can be taken from the vector P_h . However, in such applications other types of metrics need to be used during the estimation of the model. It is common to use the log loss, which is also known as the cross-entropy loss. Discrete choice models and (multinomial) logistic regression models typically use this loss function during estimation, which means that predicted probabilities from these models tend to be close to the true probabilities.

When assessing the quality of a model P the log loss can be computed, but additionally it may be insightful to compute a reliability diagram (Silva Filho et al. 2023), which visually shows the difference between predicted and empirical probabilities. Formally, for a prediction model P it should hold that $\mathbb{P}(i \text{ cars} \mid P_h(i) = z) = z$. In other words, households for which the model states “the household has 2 cars with probability 0.8” should have 2 cars in 80 percent of the cases. Models for which this property holds are called *calibrated* in the probabilistic sense, meaning that their predicted probabilities are consistent with observed outcomes without requiring additional adjustment. A reliability diagram, for which an illustrative example is shown in Figure 1, bins all predictions based on the predicted scores, and it compares the average score per bin with the empirical percentage that is seen in the data. In the figure each dot corresponds to a bin containing datapoints. Under ideal circumstances the empirical percentages are perfectly aligned with the diagonal of the diagram, but in practice this is rarely the case, meaning that models tend to under- or overestimate the actual probability. The deviation between the predicted probability and the empirical percentage can be used to compute the expected calibration error for i cars as follows:

$$\sum_{q=1, \dots, M} \frac{|\mathbb{B}_q|}{N} \mid A(\mathbb{B}_q) - \bar{A}(\mathbb{B}_q) \mid, \quad (1)$$

in which M is the number of bins, N is the total number of households, $\mathbb{B}_q$ is the set of predictions belonging to bin q , $A(\mathbb{B}_q)$ denotes the average prediction for i cars in

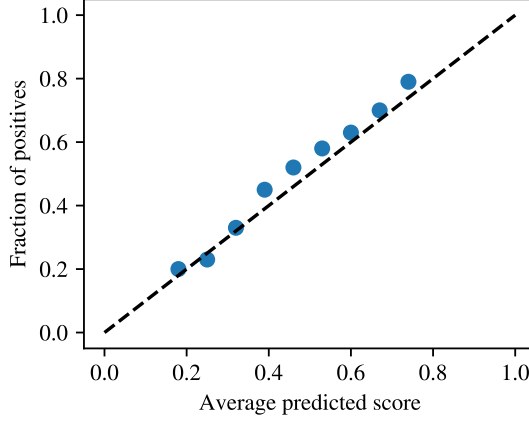


Fig. 1: Example reliability diagram for $P_h(i)$

bin $\mathbb{B}_q$, and $\bar{A}(\mathbb{B}_q)$ is the empirical percentage of households with i cars in bin $\mathbb{B}_q$. It is important to note that the expected calibration error is computed per class, but it can be generalized to multiple classes by taking the weighted sum of classwise errors. Furthermore, it is also important to note that reliability diagrams and expected calibration error evaluate the consistency between predicted and observed outcomes at the level of individual households. They therefore assess household-level calibration rather than the quality of aggregated predictions for population groups.

Several post processing methods exist to calibrate a given prediction model, which typically involves estimating an additional model which adjusts the predictions provided by the initial model. Common approaches include logistic calibration and isotonic regression, as well as algorithms which explicitly calibrate for multiple classes if the prediction model predicts for more than two classes (Silva Filho et al. 2023). All these approaches evaluate or adjust probabilities using information aggregated across the calibration dataset. Their objective is to improve the reliability of individual probability estimates, rather than to explicitly improve the accuracy of distributions predicted for predefined population groups.

2.2 Predictions for population groups

Prediction models are increasingly being used to inform policy decisions, and therefore it becomes crucial that probability estimates are well-calibrated in order to maintain trust and transparency in analyses based on these models. Many transport policy questions are formulated at the level of population groups rather than individual households. Examples include equity analyses, distributional impact assessment, and comparisons between socio-economic groups. In such applications, the primary quantity of interest is the predicted distribution for a population group. A population group $G \subseteq H$ is a group of households with the same properties. We let G_j denote

property j of group G , such that the households belonging to the group are defined by:

$$G = \left\{ h \in H \mid \bigwedge_{j=1, \dots, m} h_j = G_j \right\}. \quad (2)$$

For a given group G the predicted probability distribution over the number of cars is computed as follows:

$$P_G = \sum_{h \in G} \frac{1}{|G|} P_h, \quad (3)$$

in which the summation takes the element wise sum, such that the end result is a distribution with $n+1$ entries. In other words, the prediction is obtained by taking the mean of the predictions corresponding to households in G . The resulting prediction shows how the number of vehicles is expected to be distributed across households in the population group. This distribution is often the quantity ultimately used in policy analysis and therefore serves as the primary object of interest in the remainder of this paper.

When analyzing specific groups of the population, it is crucial that the predicted distribution P_G is close to the true distribution in this group, which is denoted by $\bar{P}_G$. For a given group G the difference between the predicted and the actual distribution can be measured by computing the Kullback–Leibler divergence (Kullback 1968):

$$D_{KL}(\bar{P}_G \parallel P_G) = \sum_{k=0, \dots, n} \bar{P}_G(k) \cdot \log \left(\frac{\bar{P}_G(k)}{P_G(k)} \right). \quad (4)$$

It is important to observe that calibration at the level of individual households, as discussed in the previous section, does not necessarily ensure that predictions at the level of population groups are also close to the true distribution. Household-level calibration assesses whether predicted probabilities are consistent with observed outcomes for individual households, while population-group-level evaluation assesses whether aggregated predictions correctly reproduce the distribution observed within a specific group. It is important to note that household-level calibration and population-group-level prediction accuracy evaluate different quantities. Standard calibration methods and reliability diagrams focus on the first objective and therefore do not explicitly evaluate the second. Consequently, a model may appear well calibrated according to household-level metrics while still producing inaccurate probability distributions for particular population groups. This distinction is sometimes overlooked in the literature, where probabilistic prediction models are typically evaluated using metrics defined at the level of individual observations.

In this section we have introduced and defined the role of population groups. However, for practical applications it is important to mention that the definition of these groups also depend on policy priorities and the availability of data. It is important that group characteristics are carefully selected, to ensure that prediction models and the studies based on these models are meaningful for policy makers. The distinction between household-level and population-group-level prediction introduced here

forms the basis for the remainder of this paper, where we study population-group-level calibration as a separate prediction objective.

3 Calibration for population-group predictions

In the previous section we distinguished between household-level calibration and population-group-level prediction accuracy. While existing calibration methods primarily focus on the former, many transport policy questions are formulated at the level of population groups. Prediction models are typically calibrated using data from the entire population, which aims to ensure reliable probability estimates at the household level. However, good household-level calibration does not necessarily imply accurate predictions for specific population groups. As a result, a model may appear well calibrated according to standard metrics while still producing biased distributions for particular groups. To address this issue, we propose a post-processing approach that explicitly targets subgroup-level calibration. As we will show in the evaluation, substantial subgroup-level prediction errors can remain even when household-level metrics indicate good performance, and the proposed approach aims to reduce such errors.

3.1 Subgroup-level calibration objective and approach

Household-level calibration evaluates whether predicted probabilities are reliable for individual households. In contrast, subgroup-level calibration evaluates whether aggregated predictions correctly reproduce the observed distribution within a population group. The objective considered in this paper is therefore to minimize the discrepancy between P_G and $\bar{P}_G$ for the groups that are relevant to a particular application. The approach that we propose explicitly calibrates at group level, in such a way that the predicted distributions P_G become closer to the true distributions $\bar{P}_G$ for each population group G . Formally, we adjust the original group-based prediction function from Equation 3 by introducing a scalar vector z^G as follows:

$$\hat{P}_G = \sum_{h \in G} \frac{1}{|G|} P_h \cdot z^G. \quad (5)$$

The vector z^G should be chosen in such a way that $D_{KL}(\bar{P}_G || \hat{P}_G)$ is minimized. The entries z_k^G can be obtained for each k by solving the following equation:

$$\left(\frac{1}{|G|} \sum_{h \in G} P_h(k) \right) \cdot z_k^G = \bar{P}_G(k), \quad (6)$$

in which $\bar{P}_G(k)$ is derived from a calibration training dataset, and $P_h(k)$ comes from the original uncalibrated prediction model. Intuitively, all predicted scores for k cars are multiplied by scalar z_k^G to make sure that the group based average equals the true empirical probability $\bar{P}_G(k)$. Because the correction is applied to predicted probabilities rather than model parameters, the approach does not depend on the underlying

Table 1: Predictions for households $h \in G$, and the group-based prediction (bottom row)

h	$P_h(0)$	$P_h(1)$	$P_h(2)$	$P_h(3)$
h_0	0.15	0.10	0.52	0.23
h_1	0.12	0.13	0.48	0.27
h_2	0.10	0.15	0.50	0.25
P_G	0.12	0.13	0.50	0.25

model structure and can be used with discrete choice models as well as machine learning models. Importantly, the correction is applied after model estimation and therefore does not require re-estimating the underlying model.

Calibration vectors z^G are computed for each group by enumerating all groups G , and for each group the computation takes place. The resulting vector is only applied to predictions for group G during prediction if $D_{KL}(\bar{P}_G \parallel \hat{P}_G) < D_{KL}(\bar{P}_G \parallel P_G)$. This validation check prevents introducing corrections that improve calibration on the calibration-training subset but do not generalize to unseen observations. The latter needs to be evaluated on a calibration test dataset, to make sure that the entire calibration step is performed on unseen data. When applying the vectors z^G at prediction time it may happen that the entries of the vector $P_h \cdot z^G$ in Equation 5 do not sum to precisely 1, which means that a normalization step is required in the actual implementation. An additional advantage of the proposed calibration method is the computational efficiency, because the scaling factors can be computed in closed form, and they are applied in a post-processing step after making the predictions. This is further illustrated in the example in the next section.

It is important to note that differences at the level of population groups can also be addressed during model estimation, for example by including interaction effects, group-specific alternative-specific constants, random coefficients, latent classes, or by estimating subgroup-specific models. Such approaches increase model flexibility and aim to capture heterogeneity directly, but typically require more complex model specifications and additional estimation effort. In contrast, the calibration method proposed in this section operates as a post-processing step applied after model estimation. This makes it complementary to model specification choices, and allows it to be applied to a wide range of existing models to improve group-level predictions without requiring changes to the model structure or re-estimation. In the next section, we demonstrate that differences at the level of population groups arise in practice across different model types, and that the proposed method can be used to reduce such errors through a model-agnostic post-processing step.

3.2 Calibration example

We illustrate the calibration approach using a simple example based on a population group G with three households and $n = 3$. The example illustrates that the method operates directly on predicted probabilities and does not require access to the internal

structure of the underlying prediction model. The original predictions from the prediction model are shown in Table 1, with the predictions P_h in the first three rows, and the prediction for the entire population group based on Equation 3 in the bottom row. Suppose that $\bar{P}_G$ is equal to $[0.13, 0.12, 0.51, 0.24]$ according to the data, then we need to compute a calibration vector z^G to make sure that the calibrated prediction $\hat{P}_G$ becomes equal to the distribution $\bar{P}_G$. The calibration vector is computed as follows:

$$z_0^G = \frac{\bar{P}_G(0)}{\frac{1}{|G|} \sum_{h \in G} P_h(0)} = \frac{0.13}{\frac{1}{3} (0.15 + 0.12 + 0.10)} = 1.054, \quad (7)$$

$$z_1^G = \frac{\bar{P}_G(1)}{\frac{1}{|G|} \sum_{h \in G} P_h(1)} = \frac{0.12}{\frac{1}{3} (0.10 + 0.13 + 0.15)} = 0.947, \quad (8)$$

$$z_2^G = \frac{\bar{P}_G(2)}{\frac{1}{|G|} \sum_{h \in G} P_h(2)} = \frac{0.51}{\frac{1}{3} (0.52 + 0.48 + 0.50)} = 1.020, \quad (9)$$

$$z_3^G = \frac{\bar{P}_G(3)}{\frac{1}{|G|} \sum_{h \in G} P_h(3)} = \frac{0.24}{\frac{1}{3} (0.23 + 0.27 + 0.25)} = 0.960. \quad (10)$$

The computed values are used to scale the scores that are predicted by the original uncalibrated model. Applying the vector $z^G = [1.054, 0.947, 1.020, 0.960]$ in Equation 5 yields the prediction $[0.13, 0.12, 0.51, 0.24]$, which is equal to the distribution that is seen in the data.

4 Experiments

In this section, we perform several computational experiments to evaluate and compare car ownership prediction models. We first discuss the underlying data and prediction models in Section 4.1. In Section 4.2, we analyze the data based on several descriptive statistics and visualizations. The remaining sections provide a series of experiments to compare prediction models and to study the effectiveness of our calibration approach. The experiments are designed to investigate three questions. First, does good household-level calibration imply accurate predictions for population groups? Second, can subgroup-level prediction errors be reduced through explicit subgroup calibration? Third, can such corrections be applied consistently across different model classes?

4.1 Data preparation and prediction models

In order to perform experiments, we create car ownership prediction models based on a large dataset containing households and their characteristics. This dataset is derived from the travel survey ‘Onderweg in Nederland’ (ODiN), conducted by Statistics Netherlands (Statistics Netherlands 2023). The ODiN dataset contains realized trips by a representative sample of the population in the Netherlands. From the dataset, we take reported trips from 2018, 2019, 2022, and 2023, which include household characteristics. We only keep the information about the household and the corresponding

Table 2: Features in the dataset

Feature	Type
Standardized income	5 discrete categories
Kids in household	binary
Number of driving licenses	1, 2, or 3+
Area density	continuous, discretized in 10 bins
Number of street parking lots	low, medium, high, very high
Household size	integer representing the number of people
ZIP code	four digits
Number of cars	0, 1, 2, or 3+

number of cars in that household. In particular, we use the standardized income, information about kids in the household, the number of driving licenses, the number of cars, the household size, the ZIP code, and the area density (the number of addresses per square kilometer). Furthermore, we augment the dataset with information about the number of street parking lots within the area where the household lives (Snelder et al. 2021). Table 2 provides a complete overview of the features. The households in the dataset may have 0, 1, 2 and 3 cars. The data preparation process yields a dataset containing 463,846 households. Households without a driving license are excluded from the analysis, as their car ownership is structurally constrained. The focus of this study is therefore on modeling variation in car ownership conditional on the presence of at least one licensed driver. We also removed households with incomplete survey data. Figure 2 provides a detailed overview of the dataset used for training the models. Each histogram displays the distribution of households by standardized income, presence of kids, number of driving licenses, area density (in bins), street parking availability, and the size distribution of population subgroups used in the experiments. Note that the area density histogram shows nearly uniform bin heights, which is a direct result of using quantile-based binning to ensure each density bin contains a similar number of households. Additional histograms describing other aspects of the data such as geographical spread are provided later in the paper for comparison.

We include models with different levels of flexibility and probabilistic properties to assess whether group-level calibration remains beneficial across model types. We use a multinomial logistic regression model (MLR), which assumes that households choose among discrete alternatives (0, 1, 2, or 3+ cars) based on a linear utility function of the observed features. The model is estimated by maximizing the log-likelihood and provides probabilistic predictions over the discrete outcomes. To explicitly account for the ordinal structure of the outcome, we also include an ordered logit model (OL). This model assumes an underlying latent variable, partitioned by thresholds into ordered categories, and produces class probabilities via cumulative logistic functions. In contrast to MLR, it enforces monotonicity across outcome levels. We also include a random forest (RF), which is an ensemble method consisting of multiple decision trees. Predictions are obtained by averaging the class predictions of the individual trees. While flexible and able to capture nonlinear relationships and interactions, random forest

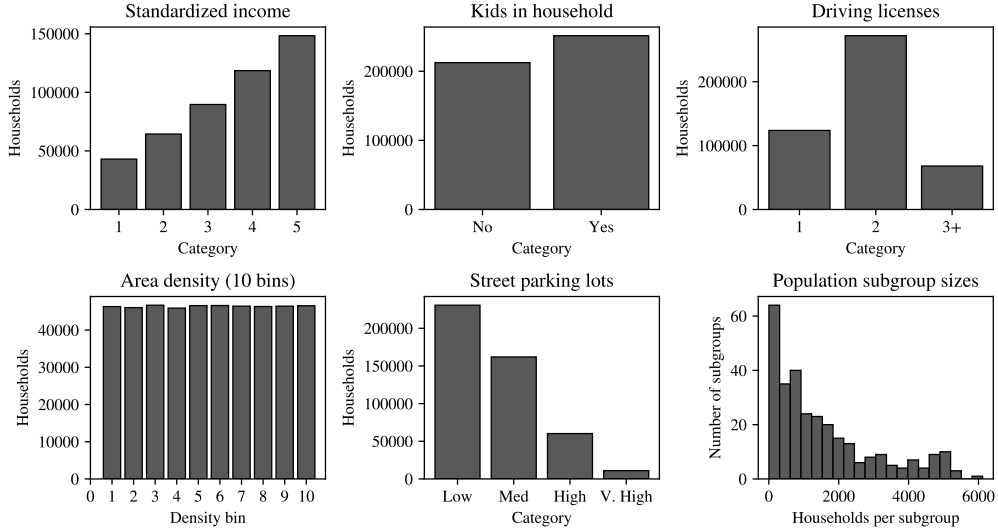


Fig. 2: Information about composition of the dataset used in the experiments

models do not directly optimize a probabilistic loss. Finally, we also include gradient boosting (GB), which builds an ensemble of decision trees sequentially, where each tree is trained to improve the predictions of the previous ensemble. The model is optimized with respect to log loss and is therefore expected to provide well-calibrated probability estimates at the household level. For each model we additionally consider two calibration approaches. First, we apply isotonic regression as a standard method to improve calibration at the household level. Second, we apply the group-based calibration method proposed in this paper, which explicitly targets the accuracy of predicted distributions at the level of population groups.

We use $\frac{1}{3}$ of the data for training the prediction models, $\frac{1}{3}$ of the data for testing, and $\frac{1}{3}$ of the data is eventually used for calibration. The data preparation took place in Python, and the prediction models are fitted using the scikit-learn package. More details about the models and parameters can be found in the appendix.¹

4.2 Motivation for subgroup-level prediction

Before evaluating prediction models, we first investigate the variability of car ownership across geographical areas and population groups. The objective is to show that many transport policy questions are inherently subgroup-oriented and that differences exist between population groups. First, we look at the geographical spread of car ownership (at the ZIP code level). Then, we investigate car ownership across various population groups defined by the univariate and bivariate combinations of variables in the dataset.

¹Source code will be available on github and on the homepage of the first author.

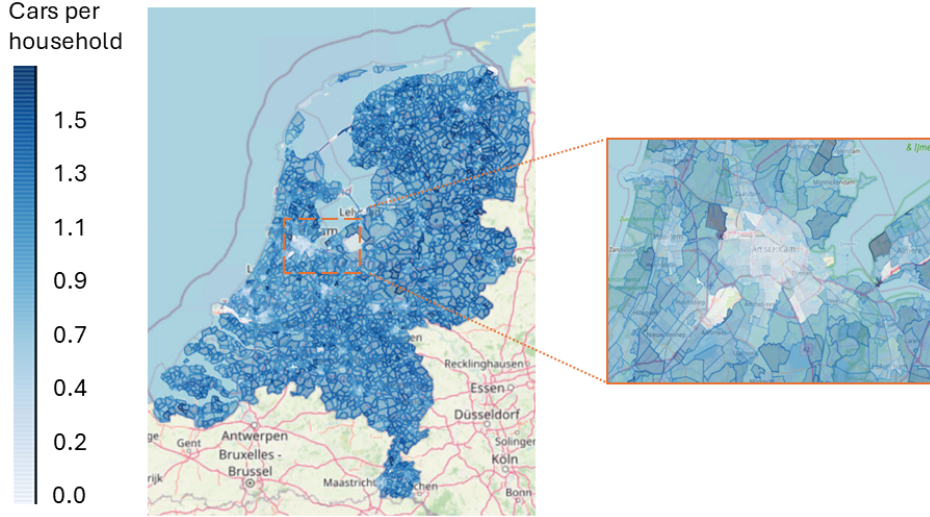


Fig. 3: Average car ownership per household in Dutch ZIP code areas

Car ownership - geographical spread

Figure 3 shows the average number of cars per household per ZIP code area, showing high variability between the different areas. There seem to be fewer cars per household in the central-western part of the Netherlands. It should be noted that this region represents the Randstad of the Netherlands, which includes the Netherlands' four biggest cities (i.e., Amsterdam, Rotterdam, The Hague, and Utrecht). Thus, there are expected to be fewer cars per household compared to more rural regions, such as the north-eastern provinces in the Netherlands. Figure 3 also zooms in to view more details in Amsterdam. Here, we can see that the number of cars per household within the city limits is very low and increases outside the city borders. These observations align with our intuitive expectations.

Car ownership distribution across population groups

We looked at the difference in distribution between the car ownership over the whole data and the car ownership distribution in population groups to show the importance of calibrating prediction models to the population groups. The analyzed population groups were defined by a univariate or bivariate combination of variables in the data.

When looking at population groups by using two or more variables, the complexity of comparing car ownership quickly increases due to the large number of groups in those cases. For example, when looking at population groups defined by only the number of driving licenses, there are only three population groups (i.e., 1, 2, and 3+). When looking at population groups defined by both the number of driving licenses in a household, as well as the household income classes (i.e., categories 1, 2, 3, 4, and 5), there are already 15 population groups to analyse.

To get a sense of the difference in car ownership distribution between various variable configurations employed to create the population groups (e.g., based on the number of driving licenses, or based on the number of driving licenses as well as the household income) and the whole distribution, the following method was applied. For each feature (combination), the KL divergence with the whole dataset distribution is computed per feature value. Then, the mean is computed over all feature values. The mean can be used to rank the feature (combinations) from most diverging to least diverging from the dataset.

For the variable configurations based on a single feature, the most diverging population groups were based on the number of driving licenses (Figure 4a). This was followed by, in descending order, the number of persons in a household (Figure 4b), the standardised income (see Figure 4c), the amount of available street parking, and whether the household had kids or not (Figure 4d). In the case of variable configurations based on two features, the combination of number of driving licenses and amount of available parking bays had the highest mean KL divergence with the whole dataset (see Figure 5 for the configuration with the second highest mean KL divergence). The figure shows, as expected, that more parking availability is related to more cars per household. It also shows that the population group with kids is more skewed to the right, having more cars. These results demonstrate that population groups can exhibit distributions that differ substantially from the overall population average. Consequently, prediction models intended for subgroup analysis should be evaluated with respect to their ability to accurately reproduce such group-specific distributions. The spatial variation in car ownership also confirms that localized policy interventions may be needed in practice, and for studying these local interventions calibrated prediction models are essential.

4.3 Household-level calibration

In this section we proceed with our computational experiments. We first evaluate the models using conventional household-level calibration metrics, namely reliability diagrams and expected calibration error. This experiment represents the type of evaluation that is commonly performed when assessing probabilistic prediction models. The reliability diagrams are depicted in Figure 6, in which the lines correspond to 0, 1, 2 and 3 cars, respectively. In the figure we compare the non-calibrated models, models calibrated with isotonic regression, and our group-based calibration approach. The corresponding expected calibration errors are provided in Table 3.

The reliability diagrams show that the multinomial logistic regression model is already well calibrated at the household level, which is consistent with its likelihood-based estimation procedure. The ordered logit and gradient boosting models also show low calibration errors, indicating that these models provide reliable probability estimates for individual households. Applying isotonic calibration or the proposed group-based calibration leads to additional improvements for several models, although the magnitude of these improvements differs across model types. Based solely on the household-level metrics shown in Figure 6 and Table 3, one could conclude that the multinomial logistic regression, ordered logit, and gradient boosting models are all suitable for probabilistic prediction tasks at the household level.

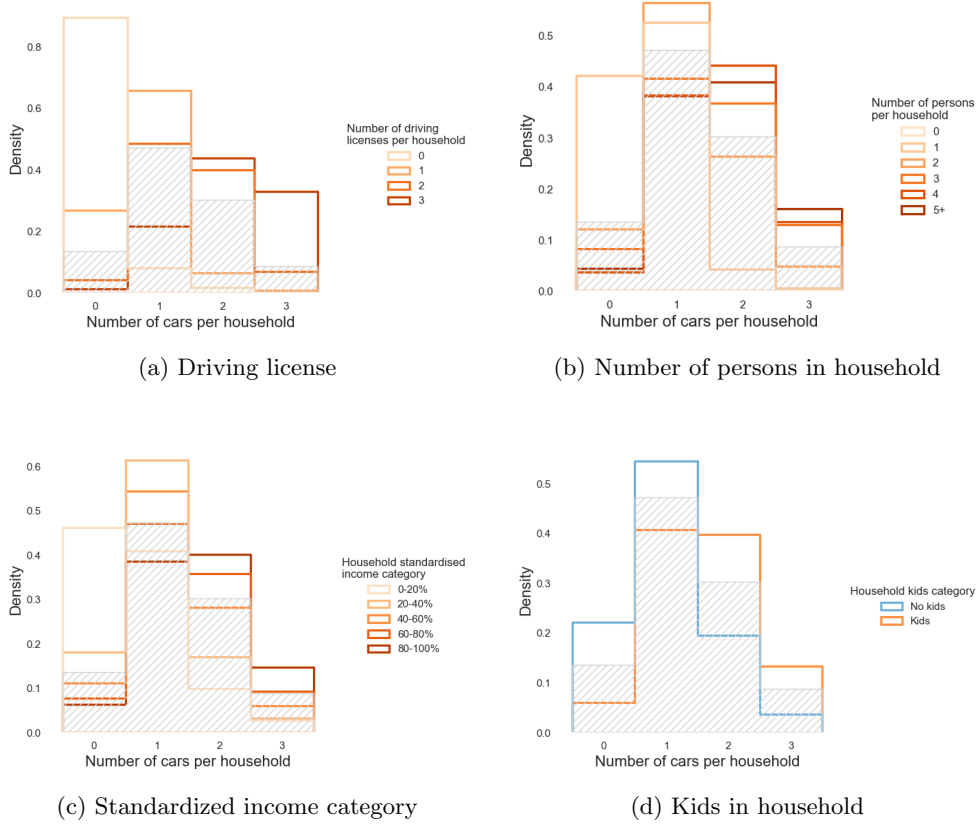


Fig. 4: Distribution of number of cars for the whole dataset (striped) versus distribution for the population groups for various features.

In contrast, the random forest shows substantially larger deviations in the reliability diagrams and the highest expected calibration error among all models considered. The model systematically under- and overestimates probabilities in several regions of the prediction space, particularly for probabilities close to zero and one. This confirms that, despite its flexibility, a random forest does not necessarily provide reliable probability estimates. Applying isotonic regression substantially reduces these errors, resulting in calibration performance that is comparable to calibrated multinomial logistic regression. Overall, this experiment highlights the importance of evaluating household-level calibration before using predicted probabilities in further analyses. At the same time, as we will show in the next section, good household-level calibration does not necessarily imply accurate predictions once probabilities are aggregated over population groups.

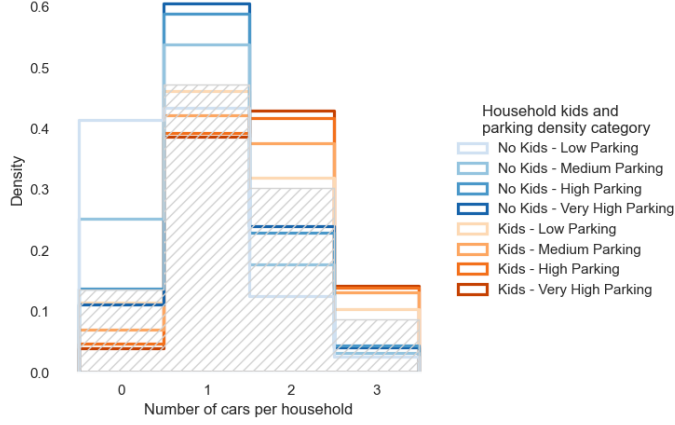


Fig. 5: Distribution of number of cars for the whole dataset (striped) versus distribution for the population groups without (blue) and with kids (orange). Darker colors reflect a higher parking availability.

Table 3: Expected calibration error and group-based KL divergence of models

Model	Expected calibration error	KL divergence for groups
MLR	0.019	0.028 ± 0.027
MLR + isotonic calibration	0.014	0.032 ± 0.035
MLR + group calibration	0.005	0.018 ± 0.025
OL	0.014	0.027 ± 0.029
OL + isotonic calibration	0.003	0.024 ± 0.030
OL + group calibration	0.004	0.018 ± 0.028
RF	0.057	0.007 ± 0.015
RF + isotonic calibration	0.004	0.016 ± 0.014
RF + group calibration	0.058	0.008 ± 0.020
GB	0.004	0.010 ± 0.013
GB + isotonic calibration	0.009	0.011 ± 0.014
GB + group calibration	0.003	0.011 ± 0.018

4.4 Analyzing predictions for population groups

We now evaluate the same models using the subgroup-level objective introduced in Section 2. Specifically, we measure the divergence between predicted and observed car ownership distributions within population groups. This experiment investigates whether conclusions obtained from household-level metrics remain valid when models are used for subgroup analysis. We evaluate this using the KL divergence defined in Equation 4, and we compute the predictions and the corresponding divergence

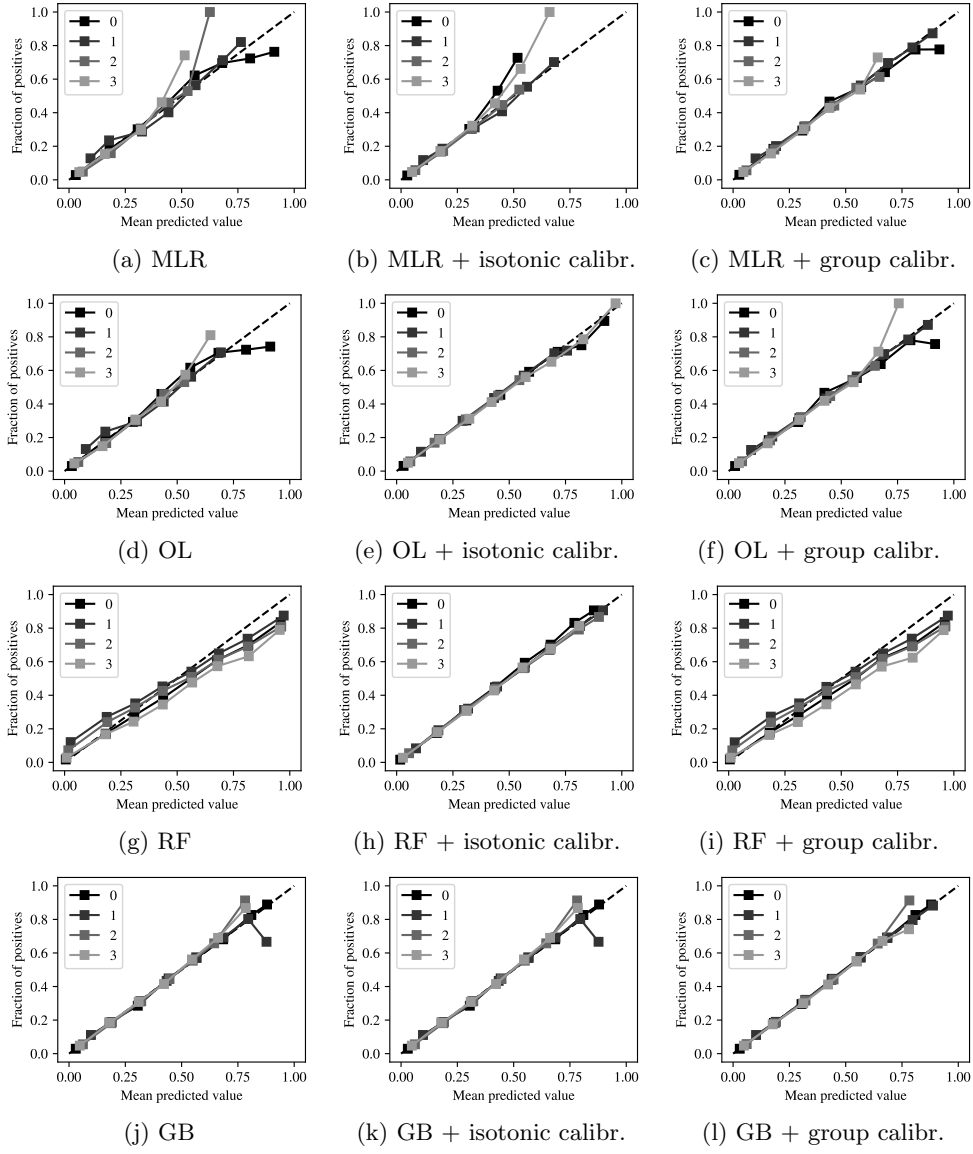


Fig. 6: Reliability diagrams for multinomial logistic regression (MLR), ordered logit (OL), random forest (RF) and gradient boosting (GB). Lines are for different numbers of cars at household level.

separately for each population group G , as shown in the boxplot in Figure 7. The means and standard deviations are provided in Table 3.

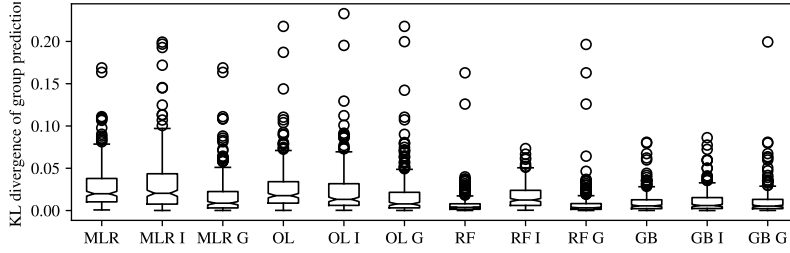


Fig. 7: KL divergence of group based predictions

Across all models, we observe that deviations between predicted and observed distributions can occur at the level of population groups. This is an important observation because several of the models considered in the previous section have already shown good calibration properties at the household level. The results therefore confirm that household-level calibration and population-group-level prediction accuracy are distinct objectives. Models that appear reliable according to conventional household-level metrics may still produce biased distributions when predictions are aggregated over population groups.

The proposed group-based calibration approach directly targets this second objective by adjusting predictions to better match observed group-level distributions. For the multinomial logistic regression and ordered logit models, this leads to a clear reduction in the average KL divergence. Furthermore, the group-calibrated variants consistently outperform the corresponding isotonic calibration variants, indicating that improving household-level calibration does not necessarily improve subgroup-level predictions.

Gradient boosting provides an especially interesting example. In the previous section, the model already achieved good calibration performance at the household level. Nevertheless, Table 3 shows that deviations remain when predictions are evaluated at the level of population groups. Although the improvement obtained through group-based calibration is modest, this result demonstrates that subgroup-level discrepancies may persist even for models that optimize probabilistic loss functions and provide well-calibrated household-level predictions. This supports the main premise of the paper that subgroup-level calibration should be considered a separate prediction objective.

The random forest shows the opposite behavior. Despite performing relatively well according to the group-level KL divergence metric, the model showed substantially poorer calibration at the household level. Applying isotonic regression largely resolves these household-level calibration errors, but at the same time leads to worse performance at the level of population groups. This illustrates that improving one objective does not necessarily improve the other. Taken together, the random forest and gradient boosting results demonstrate that conclusions about model quality depend strongly on the level at which predictions are evaluated.

Overall, these findings indicate that conventional household-level evaluation is insufficient when prediction models are used for subgroup analysis. Models should therefore be evaluated at the level at which they are ultimately used. When the goal

is to analyze car ownership distributions for specific population groups, subgroup-level metrics provide information that is not captured by household-level calibration measures alone.

Until now the experiments provide three main insights. First, good household-level calibration does not necessarily imply accurate predictions for population groups. Second, household-level calibration and subgroup-level prediction accuracy should be seen as distinct evaluation objectives. Third, subgroup-level discrepancies arise across different models, including models that already provide well-calibrated probabilities at the household level. These observations motivate the need for dedicated subgroup-level evaluation metrics and, when population-group analysis is the primary application, calibration methods that explicitly target group-level prediction quality.

4.5 Practical implications for population-group analysis

In the previous sections we evaluated prediction models using household-level and subgroup-level metrics. In this final experiment, we consider a tangible question that may arise in transport planning: how many households within a population group are expected to own 0, 1, 2, or 3 cars? This type of question is directly relevant for policy analysis and equity assessment, where decisions are often based on aggregate outcomes for specific groups rather than individual households. The prediction models are used to assess this separately for each value, and the absolute prediction errors are visualized in Figure 8. The corresponding means and standard deviations are provided in Table 4. The results show that a multinomial logistic regression model calibrated with our approach consistently outperforms the other variants of the multinomial logistic regression model, providing more accurate predictions. For the multinomial logistic regression and ordered logit models, the proposed calibration approach substantially reduces the absolute prediction errors. For the tree-based models, the impact is smaller and depends on the underlying properties of the model. This highlights the importance of calibration in this setting, and it shows that policy makers need to be cautious when applying existing regression models to population groups. If not applied carefully, models can provide inaccurate results which may affect policy decisions for such groups without noticing.

While the aggregate metrics reported in the previous experiments provide an overall assessment of model quality, they do not reveal which specific population groups are difficult to predict. To gain further insight, we therefore analyze the groups with the largest absolute prediction errors. We consider the multinomial logistic regression model with group-based calibration (MLR G), and for each car ownership category we identify the five population groups with the highest absolute error. The corresponding group characteristics, together with the observed and predicted numbers of households, are reported in Table 5. For interpreting the feature values, it may be helpful to consult Table 2.

The table shows that substantial errors may still occur for particular population groups, even when the overall performance of the model is strong. For example, the group consisting of households with income class 2, no kids, one driving license, and a medium-density residential environment exhibits a large error for both the 0-car and 1-car categories. Such deviations can be problematic when model outputs are used to

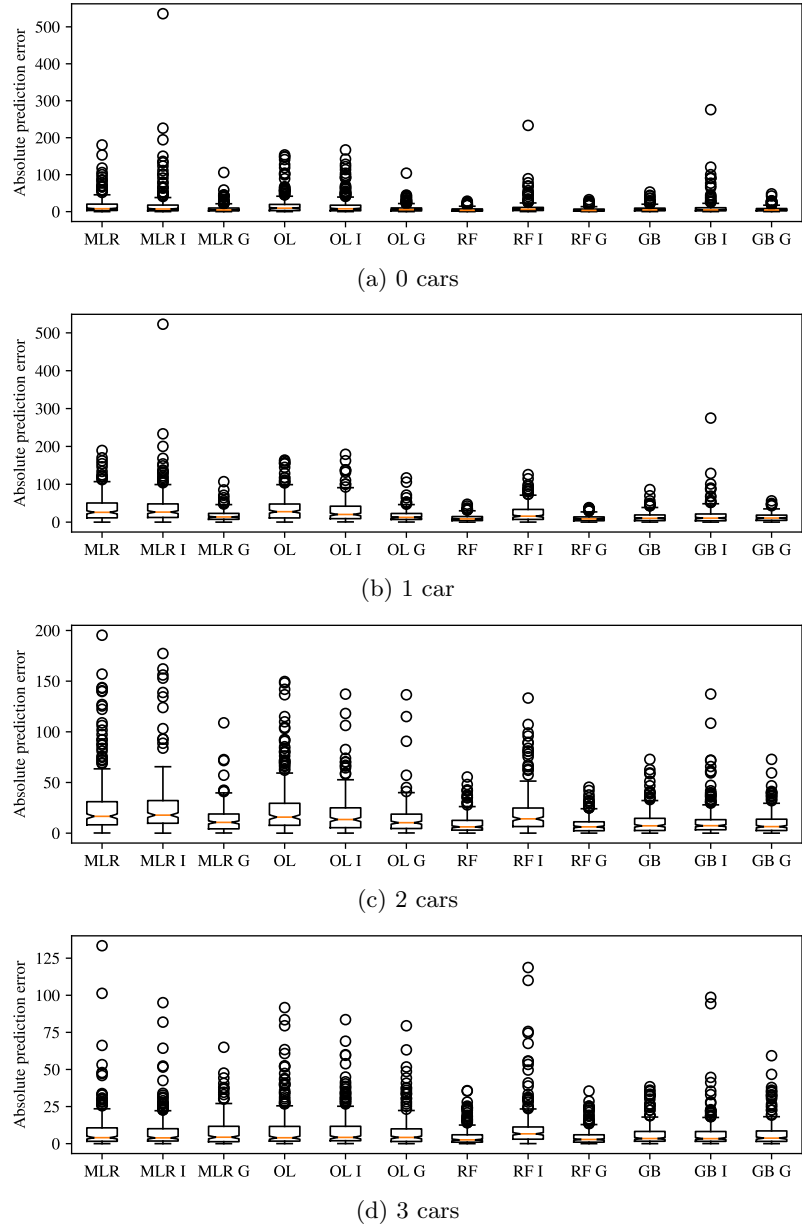


Fig. 8: Absolute errors on the number of households with 0, 1, 2 and 3 cars in a population group

support policy decisions that target specific population groups. Rather than relying solely on aggregate performance measures, these groups require additional attention when interpreting model results.

Table 4: Absolute errors

Model	0 cars	1 car	2 cars	3 cars
MLR	16.35 ± 24.22	35.85 ± 33.55	27.13 ± 31.99	9.04 ± 14.42
MLR + isotonic calibr.	19.32 ± 45.02	39.13 ± 47.86	26.57 ± 28.59	8.41 ± 12.13
MLR + group calibr.	8.33 ± 11.25	17.13 ± 14.57	13.74 ± 13.48	8.02 ± 9.86
OL	17.59 ± 26.69	35.49 ± 32.91	25.36 ± 27.85	9.95 ± 14.74
OL + isotonic calibr.	16.51 ± 25.59	30.94 ± 30.07	18.69 ± 19.68	9.47 ± 12.75
OL + group calibr.	8.20 ± 10.42	16.97 ± 15.08	14.24 ± 15.69	7.93 ± 10.89
RF	5.23 ± 5.23	10.66 ± 9.00	9.34 ± 9.53	5.01 ± 6.46
RF + isotonic calibr.	10.76 ± 19.22	23.35 ± 22.20	20.35 ± 21.04	10.91 ± 15.24
RF + group calibr.	5.08 ± 5.59	9.77 ± 8.05	8.32 ± 7.95	4.74 ± 5.77
GB	7.33 ± 7.97	13.23 ± 12.69	10.78 ± 11.91	6.45 ± 7.69
GB + isotonic calibr.	10.76 ± 23.99	16.84 ± 24.73	12.23 ± 16.52	6.84 ± 10.98
GB + group calibr.	6.58 ± 6.91	12.77 ± 10.39	10.02 ± 10.79	6.82 ± 8.75

More generally, this analysis illustrates why household-level and aggregate evaluation metrics alone are insufficient for subgroup-oriented policy applications. Metrics such as log loss, expected calibration error, or average KL divergence provide valuable information about overall predictive performance, but they do not reveal how errors are distributed across individual population groups. Inspecting subgroup-level prediction errors therefore remains essential when prediction models are used for e.g. equity assessment and distributional analysis.

5 Related work

In this section we provide an overview of related work and we summarize the similarities and differences with our work. Several people have studied the link between machine learning methods and choice modeling, reflecting that there is increasing interest in both traditional and data-driven approaches for transport modeling. The characteristics of logit models and other machine learning classifiers has been studied in previous work on choice modeling in transport, mainly in the context of mode choice modeling and car ownership transactions (Hillel et al. 2021; Ali et al. 2023). Typically it is observed that classifiers such as random forests have higher predictive power, but existing research also notes that such models do not always produce reasonable behavioural outputs such as elasticities and the value of time (Zhao et al. 2020). For choice modeling using neural networks (van Cranenburgh and Alwosheel 2019) it has been shown that quantities such as value of time can be revealed with such more complex classifiers (van Cranenburgh and Kouwenhoven 2021). Our work also studies the strenghts and limitations of machine learning models such as random forests, but by focusing on population groups we approach it from a different angle.

In addition to household characteristics, which we studied in this paper, there are other factors influencing car ownership decisions. In particular, the birth of a baby, a marital status change or income change can lead to changes in car ownership, sometimes later in time (Dargay and Gately 1999). Other factors are changes in travel

Table 5: Population groups with the highest absolute prediction errors for model MLR G

Number of cars	Income	Kids	Driving licenses	Density	True value	Predicted
0 cars	2	0	1	6	185	395
	5	0	1	7	288	203
	2	0	1	5	190	274
	1	0	3	9	96	25
	1	0	1	8	1467	1398
1 car	2	0	1	6	1008	799
	5	1	2	5	1315	1482
	5	1	2	1	652	803
	4	0	2	3	1356	1229
	5	1	2	8	1729	1620
2 cars	5	1	2	5	2297	2089
	5	1	2	1	1777	1597
	5	1	2	8	1345	1473
	4	0	1	4	40	123
	4	0	1	0	90	167
3 cars	5	1	3	0	1601	1695
	4	1	3	0	841	768
	4	1	2	3	240	180
	5	1	2	7	223	279
	1	1	3	0	109	52

times that people experience (Ding et al. 2017). Modeling such car ownership decisions typically require temporal data which capture the characteristics of households over a longer period of time (Dargay 2001), which may be modeled by using e.g. dynamic Bayesian networks (Yang et al. 2023). Lag effects and factors triggering changes in car ownership are not captured in our work, and can be seen as an interesting angle for next step since triggers may also differ depending on the population group that is considered.

In a more general context our work is also related to the analysis of the effects on various socioeconomic groups in cities, also known as distribution effects (Iglesias et al. 2019). This not only involves the analysis of effects on groups, but also the process of identifying the groups that are relevant to analyze in a given situation (Biscaro Uliana et al. 2024). Since our work is a general methodology to obtain improved predictions for prediction tasks based on population groups, we expect that our work may be generalized to other types of prediction tasks in an urban context, contributing to improved assessment of distribution effects.

In the field of machine learning there is a large body of work on probability calibration (Silva Filho et al. 2023), which ensures that outputs provided by a prediction model are aligned with the true probabilities. This is crucial in critical applications,

and becoming even more important in the development of trustworthy artificial intelligence systems (Kaur et al. 2022). Most calibration methods focus on classifiers for prediction tasks with only two classes (e.g., 0 and 1), which can become problematic if there are more than two classes. Recent methods such as Dirichlet calibration (Kull et al. 2019) have emerged as a methodology to perform calibration in a multi-class context. Our work also advances calibration methodologies, but instead of focusing on the number of classes we study groups of datapoints representing population groups. We position our work more in a broad landscape, focusing on the question how to integrate calibration techniques and group-based analysis in transport modeling.

6 Conclusion

Predicting car ownership is an important part of transport planning. While prediction models are commonly evaluated using household-level metrics such as log loss, reliability diagrams, and expected calibration error, many practical applications rely on predictions aggregated over population groups. In this paper, we demonstrated that good household-level calibration does not necessarily imply accurate predictions at the level of population groups. Models that provide reliable probability estimates for individual households can still produce biased distributions when predictions are aggregated over groups.

This observation motivates a distinction between household-level calibration and subgroup-level calibration. The former evaluates whether predicted probabilities are consistent with observed outcomes for individual households, whereas the latter evaluates whether aggregated predictions reproduce the observed distributions within population groups. Our results show that these objectives are related but not equivalent, indicating that evaluation metrics should be aligned with the level at which model outputs are ultimately used.

To address subgroup-level prediction errors, we proposed a model-agnostic calibration method that operates as a post-processing step on predicted probabilities. Because the method does not depend on the internal structure of the prediction model, it can be applied to different model classes without requiring re-estimation. In a case study based on a large dataset of Dutch households, we evaluated multinomial logit, ordered logit, random forest, and gradient boosting models. The results showed that subgroup-level discrepancies occur across different model types, including models that already provide well-calibrated probabilities at household level. The proposed approach reduced subgroup-level prediction errors for multiple models while maintaining the flexibility to work across different modeling paradigms.

From a practical perspective, the results highlight the importance of evaluating prediction models at the level at which they are used. This is particularly relevant for transport applications involving population-group analysis, equity assessment, and distributional impact evaluation, where conclusions are often drawn from aggregated predictions rather than individual household outcomes. Household-level calibration metrics remain valuable, but they should not be used as the only criterion when model outputs are intended to support subgroup-oriented policy decisions.

In future work we can investigate the interaction between subgroup-level calibration and alternative modeling strategies, such as richer model specifications, interaction effects, mixed logit models, and subgroup-specific constants. In addition, the proposed framework may be applied to other transport modeling problems and more broadly to prediction tasks in which decisions depend on accurate distributions for specific population groups.

Statements and declarations

Competing interests: the authors declare no competing interests.

References

- Ali, A., Kalatian, A., Choudhury, C.F.: Comparing and contrasting choice model and machine learning techniques in the context of vehicle ownership decisions. *Transportation Research Part A: Policy and Practice* **173**, 103727 (2023)
- Biscaro Uliana, M., Vonk Noordegraaf, D., Wilmink, I., Vonk, T.: Defining new population groups for mobility studies. Available at SSRN 4954126 (2024)
- Cheng, L., Chen, X., De Vos, J., Lai, X., Witlox, F.: Applying a random forest method approach to model travel mode choice behavior. *Travel behaviour and society* **14**, 1–10 (2019)
- Dargay, J.M.: The effect of income on car ownership: evidence of asymmetry. *Transportation Research Part A: Policy and Practice* **35**(9), 807–821 (2001)
- Dargay, J., Gately, D.: Income’s effect on car and vehicle ownership, worldwide: 1960–2015. *Transportation Research Part A: Policy and Practice* **33**(2), 101–138 (1999)
- Ding, C., Wang, D., Liu, C., Zhang, Y., Yang, J.: Exploring the influence of built environment on travel mode choice considering the mediating effects of car ownership and travel distance. *Transportation Research Part A: Policy and Practice* **100**, 65–80 (2017)
- Hillel, T., Bierlaire, M., Elshafie, M.Z., Jin, Y.: A systematic review of machine learning classification methodologies for modelling passenger mode choice. *Journal of choice modelling* **38**, 100221 (2021)
- Iglesias, V., Giraldez, F., Tiznado-Aitken, I., Muñoz, J.C.: How uneven is the urban mobility playing field? inequalities among socioeconomic groups in santiago de chile. *Transportation research record* **2673**(11), 59–70 (2019)
- Jorritsma, P., Arendsen, K., Faber, R.: Achtergrondrapport - Autoluw Beleid Gemeenten: Doelen, Effecten, en Rollen (2023). Kennisinstituut voor Mobiliteitsbeleid (KiM)

- Kashifi, M.T., Jamal, A., Kashefi, M.S., Almoshaogeh, M., Rahman, S.M.: Predicting the travel mode choice with interpretable machine learning techniques: A comparative study. *Travel Behaviour and Society* **29**, 279–296 (2022)
- Kull, M., Perello Nieto, M., Kängsepp, M., Silva Filho, T., Song, H., Flach, P.: Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. *Advances in neural information processing systems* **32** (2019)
- Kullback, S.: *Information theory and statistics*. New York: Dover (1968)
- Kaur, D., Uslu, S., Rittichier, K.J., Durrezi, A.: Trustworthy artificial intelligence: a review. *ACM computing surveys (CSUR)* **55**(2), 1–38 (2022)
- Li, B.: The multinomial logit model revisited: A semi-parametric approach in discrete choice analysis. *Transportation Research Part B: Methodological* **45**(3), 461–473 (2011)
- Mao, X., Huang, X., Song, Y., Zhu, Y., Tan, Q.: Response to urban land scarcity in growing megacities: Urban containment or inter-city connection? *Cities* **96**, 102399 (2020)
- Paredes, M., Hemberg, E., O'Reilly, U.-M., Zegras, C.: Machine learning or discrete choice models for car ownership demand estimation and prediction? In: *2017 5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)*, pp. 780–785 (2017). IEEE
- Rainio, O., Teuho, J., Klén, R.: Evaluation metrics and statistical tests for machine learning. *Scientific Reports* **14**(1), 6086 (2024)
- Snelder, M., Araghi, Y., Ashari, B., Charoniti, E., Klunder, G., Sterkenburg, R., Tuin, M., Spruijtenburg, D., Kochan, B., Bellemans, T., et al.: *Rapport A: Methode Urban Tools Next II-toelichting op gekozen aanpak voor parkeren, ketens en hubs, nieuwe mobiliteitsconcepten*, Den Haag (The Netherlands) (2021)
- Silva Filho, T., Song, H., Perello-Nieto, M., Santos-Rodriguez, R., Kull, M., Flach, P.: Classifier calibration: a survey on how to assess and improve predicted class probabilities. *Machine Learning* **112**(9), 3211–3260 (2023)
- Statistics Netherlands: *Onderweg in Nederland (ODiN) 2023 - Onderzoeksbeschrijving* (2023)
- Cranenburgh, S., Alwosheel, A.: An artificial neural network based approach to investigate travellers' decision rules. *Transportation Research Part C: Emerging Technologies* **98**, 152–166 (2019)
- Cranenburgh, S., Kouwenhoven, M.: An artificial neural network based method to uncover the value-of-travel-time distribution. *Transportation* **48**(5), 2545–2583

(2021)

Yang, Y., Rasouli, S., Liao, F.: Effects of life events and attitudes on vehicle transactions: A dynamic bayesian network approach. *Transportation research part C: emerging technologies* **147**, 103988 (2023)

Zhao, X., Yan, X., Yu, A., Van Hentenryck, P.: Prediction and behavioral analysis of travel mode choice: A comparison of machine learning and logit models. *Travel behaviour and society* **20**, 22–35 (2020)

A Details of models and parameters

All models predict household car ownership classes $\{0, 1, 2, 3\}$ using the same base features. We use a 33.3%/33.3%/33.3% split for training, testing, and calibration. The algorithms have been implemented in Python using scikit-learn.

Multinomial Logistic Regression (MLR)

The baseline model is multinomial logistic regression, estimated by maximizing the multinomial likelihood over the four ownership classes. This provides class probabilities directly and serves as reference for the other approaches. Relevant settings are a multinomial formulation with an `lbfgs` optimizer and a sufficiently large iteration cap (`max_iter=500`) to ensure convergence.

MLR with Isotonic Calibration (MLR I)

This variant applies post-hoc isotonic calibration to the MLR probabilities on a dedicated calibration split. The purpose is to improve probabilistic calibration while leaving the ranking structure from the base classifier unchanged. The key calibration setting is isotonic regression (`method=isotonic`).

MLR with Group Calibration (MLR G)

This approach starts from the MLR predictions and applies subgroup-specific multiplicative scaling of class probabilities. Calibration factors are estimated per subgroup and retained only when they improve hold-out KL divergence, which limits overfitting and keeps adjustments conservative.

Ordered Logit (OL)

The ordered logit model treats car ownership as an ordinal outcome and estimates cumulative logits under the proportional-odds assumption. This leverages the natural class ordering and can improve stability for adjacent classes. The relevant setting is a logit link with BFGS-based optimization.

Ordered Logit with Isotonic Calibration (OL I)

This model calibrates the ordered-logit output via isotonic regression on cumulative probabilities (one calibrator per threshold), after which calibrated cumulative values are mapped back to class probabilities. The key settings are bounded isotonic calibration ($y \in [0, 1]$) and clipping for out-of-range inputs.

Ordered Logit with Group Calibration (OL G)

Analogous to MLR G, this variant applies subgroup-level multiplicative probability scaling to ordered-logit predictions. As above, subgroup corrections are accepted only when they improve hold-out KL divergence.

Random Forest (RF)

The random forest model captures nonlinearities and interactions through an ensemble of decision trees trained on bootstrap samples. The main capacity control is the number of trees (`n_estimators=100`), while tree-depth and split parameters are left at standard defaults.

Random Forest with Isotonic Calibration (RF I)

This variant applies isotonic post-calibration to RF class probabilities on the calibration split. The intent is to correct probability miscalibration without changing the underlying forest structure. The relevant calibration choice is `method=isotonic`.

Random Forest with Group Calibration (RF G)

This model applies the same subgroup-specific multiplicative calibration strategy to RF predictions, with subgroup corrections kept only when hold-out KL divergence improves.

Gradient Boosting (GB)

Gradient boosting is used as a sequential tree ensemble that refines predictions stage by stage. Relevant capacity/learning settings are the number of boosting stages (`n_estimators=100`) and learning rate (0.1), with standard tree depth settings.

Gradient Boosting with Isotonic Calibration (GB I)

This variant post-calibrates GB probabilities by isotonic regression on the calibration split. As for the other calibrated variants, the main setting is `method=isotonic`.

Gradient Boosting with Group Calibration (GB G)

This variant adds subgroup-specific multiplicative probability scaling on top of GB predictions and retains corrections only when hold-out KL divergence improves.

B Multinomial logit parameters

Table 6 reports the estimated parameters of the multinomial logistic regression (MLR) model. The model predicts the number of cars owned by a household (0, 1, 2, and 3 cars) using standardized income, presence of children, number of driving licenses, area density, street parking availability, and household size. The reported coefficients correspond to the four outcome classes used in the analysis. The estimated coefficients are consistent with intuition. Higher area density is associated with a higher probability of owning fewer cars, whereas households with more driving licenses, larger household size, and higher income are more likely to own multiple cars.

Table 6: Estimated MLR parameters

Variable	0 cars	1 car	2 cars	3 cars
Intercept	3.959	3.059	-1.076	-5.942
Income	-0.468	-0.133	0.219	0.381
Kids	-0.335	-0.061	0.259	0.137
Driving licenses	-1.507	-0.604	0.466	1.645
Density	4.552	0.758	-2.220	-3.089
Street parking	-0.315	0.012	0.096	0.207
Household size	-0.244	-0.034	0.111	0.167